Canonicalizing Open Knowledge Bases using Embeddings and Side Information

Shikhar Vashishth, Prince Jain, Partha Talukdar
Indian Institute of Science, India
Knowledge Graphs

- Knowledge in graph form
- **Nodes** represent entities
- **Edges** represent relationships b/w entities
- **Examples**: Freebase, Wikidata ...

*Figure source: Mining Knowledge Graphs from Text, WSDM ‘18 tutorial*
What are Open KGs?

- KGs with entities and relations **not restricted to a defined set**.

- **Construction**: Automatically extracting *(noun-phrase, relation-phrase, noun-phrase)* from unstructured text.
  - *Obama was the President of US.* → *(Obama, was president of, US)*
  - Examples: TextRunner, ReVerb, Ollie etc.

- **Use cases:**
  - Extract knowledge from a new domains without supervision.
Challenges with Open KG

- **Problem:** May store redundant and ambiguous facts
  - \((\text{Barack Obama, was president of, US})\)
  - \((\text{Obama, born in, Honolulu})\)

- Querying for “Barack Obama” will not return all extracted facts.

- **Solution:** Need to Canonicalize Open KGs
### Canonicalization

#### Noun Phrases
- Barack Obama
- Obama
- George Bush
- New York City
- NYC

#### Relation phrases:
- born_in
- took_birth_in
- is_employed_in
- works_for
- capital_of
Previous works

- **RESOLVER** system [Yates, 2009] uses string similarity based features to cluster phrases in TextRunner.

- [Galárraga, 2014] perform noun phrase canonicalization by clustering over manually-defined feature spaces which is followed by relation phrase canonicalization using AMIE [Galárraga, 2013]
Issues

- **Surface form not sufficient** for disambiguation
  - E.g. (US, America)

- **Manual feature engineering** is expensive and often sub-optimal

- **Sequentially canonicalizing** of noun and relation phrases can lead to error propagation
Contributions

- We propose CESI, a novel method for canonicalizing Open KBs using learned embeddings.

- CESI jointly canonicalize both noun phrase (NP) and relation phrase using relevant side information.

- We build a new data, ReVerb45K which has 20x more NPs than previous biggest dataset for the task.
CESI Overview

1. **Side Information Acquisition:**
   - Gathers various noun and relation phrase side information

2. **Embeddings Noun and relation phrases:**
   - Learns specialized vector embeddings

3. **Clustering Embeddings and Canonicalization:**
   - Clusters embeddings based on distance
   - Assigns a representative to each noun and relation cluster
Side Information Acquisition

- Involves identifying equivalence relations of form:
  - \( e_1 \equiv e_2 \) and \( r_1 \equiv r_2 \)

- **Entity Linking:**
  - Identify entity mention and link to KBs like Wikipedia
  - US \( \rightarrow \) United_States, America \( \rightarrow \) United_States

- **Paraphrase database (PPDB):**
  - Large collection of paraphrases in English
  - management \( \equiv \) administration, head of \( \equiv \) chief of
Side Information Acquisition

- **WordNet with Word-sense disambiguation:**
  - Identify synsets of NPs
  - picture ≡ image, plant ≡ industry

- **IDF Token Overlap:**
  - NPs and relations sharing infrequent terms
  - Warren Buffett ≡ Mr. Buffett, Mr. Gates ≡ Bill Gates

- Used 9 types of side info, refer paper for more.

- Side information used as **soft constraints**
Embeddings Noun and Relation phrases

- Several KG embedding algorithms available, we use of **HolE (Holographic Embeddings)**

- HolE assigns a **score** $\eta$ to each triple $(v, r, v')$ in KB:

  $$\eta = e_r^T (e_v \ast e_{v'})$$

- Learns embedding by **optimizing**:

  $$\sum_{i \in D_+} \sum_{j \in D_-} \max(0, \gamma + \sigma(\eta_j) - \sigma(\eta_i))$$
CESI Optimization Objective

\[
\min_{\Theta} \lambda_{\text{str}} \sum_{i \in D} \sum_{j \in D_+} \max(0, \gamma \sigma(\eta_j) - \sigma(\eta_i))
\]

\[
\sum_{\theta \in \mathcal{C}_{\text{ent}}} \frac{\lambda_{\text{ent}, \theta}}{|\mathcal{I}_{\text{ent}, \theta}|} \sum_{v, v' \in \mathcal{I}_{\text{ent}, \theta}} \|e_v - e_{v'}\|^2
\]

\[
\sum_{\phi \in \mathcal{C}_{\text{rel}}} \frac{\lambda_{\text{rel}, \phi}}{|\mathcal{I}_{\text{rel}, \phi}|} \sum_{u, u' \in \mathcal{I}_{\text{rel}, \phi}} \|r_u - r_{u'}\|^2
\]

\[
\lambda_{\text{reg}} \left( \sum_{v \in V} \|e_v\|^2 \sum_{r \in R} \|e_r\|^2 \right).
\]

HoLE Objective

Noun phrase
Side Information

Relation phrase
Side Information

Regularization

Optimized using SGD
CESI Architecture
Experiments
Evaluation Metrics

- **Macro:**
  - Fraction of pure clusters
  - Precision = 2/3

- **Micro:**
  - Purity of clusters
  - Precision = 6/7

- **Pairwise:**
  - Ratio of hits to all possible pairs
  - Precision = 4/6
NP Canonicalization

Cesi out-performs others in noun phrase canonicalization.
Effect of Side Information

Side information improves performance

- CESI
- w/o Entity Linking (EL)
- w/o EL, WordNet (WN)
- w/o EL, WN, Morph (M)
- w/o EL, WN, M, PPDB

Summary:
- Macro F1: CESI > w/o EL, WN, M, PPDB
- Micro F1: CESI > w/o EL, WN, M, PPDB
- Pairwise F1: CESI > w/o EL, WN, M, PPDB
## Relation Canonicalization

<table>
<thead>
<tr>
<th></th>
<th>Macro Precision</th>
<th>Micro Precision</th>
<th>Pairwise Precision</th>
<th>Induced Relation Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base Dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMIE</td>
<td>42.8</td>
<td>63.6</td>
<td>43.0</td>
<td>7</td>
</tr>
<tr>
<td>CESI</td>
<td><strong>88.0</strong></td>
<td><strong>93.1</strong></td>
<td><strong>88.1</strong></td>
<td><strong>210</strong></td>
</tr>
<tr>
<td><strong>Ambiguous Dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMIE</td>
<td>55.8</td>
<td>64.6</td>
<td>23.4</td>
<td>46</td>
</tr>
<tr>
<td>CESI</td>
<td><strong>76.0</strong></td>
<td><strong>91.9</strong></td>
<td><strong>80.9</strong></td>
<td><strong>952</strong></td>
</tr>
<tr>
<td><strong>ReVerb45K</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMIE</td>
<td>69.3</td>
<td>84.2</td>
<td>66.2</td>
<td>51</td>
</tr>
<tr>
<td>CESI</td>
<td><strong>77.3</strong></td>
<td><strong>87.8</strong></td>
<td><strong>72.6</strong></td>
<td><strong>2116</strong></td>
</tr>
</tbody>
</table>

CESI produces more and better relation canonicalized clusters.
Qualitative Evaluation (t-sne)
Conclusion

- Canonicalization is necessary for Open KG
- Existing approaches are based on manually feature engineering which can be sub-optimal
- CESI, presents an embedding based joint noun and relation phrase canonicalization
  - Utilizes several types of side information
  - Obtains state-of-the-art results for the problem
Questions?

Source code and data are available
github.com/malllabiisc/cesi

Contact email:
shikhar@iisc.ac.in
• References:
    https://dl.acm.org/citation.cfm?id=2662073